Multimodal Variational Autoencoders
Achint Kumar
Desiderata

- JMVAE, TELBO, MFM
- MVAE, Wu and Goodman 2018
- MMVAE, Shi et. al. 2019
- MoPoE, MEME

Ah, but a man's reach should exceed his grasp, Or what's a heaven for?
~ Robert Browning
Multimodal data representation

- Humans receive sensory data from multiple modalities (sight, touch, smell, etc).
- The brain binds the data together to form a unified representation of the object.
- We model this “unified representation” through latent space of VAE.
Example: Multimodal mouse vocalization

- $X_1$: Mouse vocalization audioclips
- $Z_1$(hypothesis): Parameters of mouse vocal cords (pressure, length), etc

- $X_2$: Neural recordings
- $Z_2$(hypothesis): Cluster of neurons corresponding to some frequency range

- Dream: Identify cluster of neurons that correspond to vocalization in certain frequency range
Unimodal VAE

$$\text{ELBO} = -D_{KL}(q_\phi(z|x) \| p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}[\ln(p_\theta(x|z))]$$
MVAE: Architecture

- Product of experts: Each expert has power to veto

\[ q_\phi(z|x_1, x_2) = p(z) \prod_{i=1}^{2} q_{\phi_i}(z|x_i) \]

- Missing expert: \( q_{\phi_i}(z|x_i) = 1 \)

\[ \mu = \left( \sum_i \mu_i T_i \right) \left( \sum_i T_i \right)^{-1} \]

\[ \sigma^2 = \left( \sum_i T_i \right)^{-1} \]

Wu and Goodman, 2018
MVAE: Loss function

- Learning POE Gaussian doesn’t specify individual Gaussian.

- If we learn ELBO separately then we won’t learn the relationship between modalities

- So, we add to composite ELBO the individual ELBO to get full loss function

\[
\mathcal{L} = \text{ELBO}(x_1, x_2) + \text{ELBO}(x_1) + \text{ELBO}(x_2)
\]

\[
\text{ELBO}(x_1, x_2) = \mathbb{E}_{q_{\phi}(z|x_1, x_2)} \left[ \log(p_{\theta}(x_1, x_2|z)) \right] - \text{KL}[q_{\phi}(z|x_1, x_2), p(z)]
\]
MVAE: Strengths and Weaknesses

• The loss function is not a valid lower bound on the joint log-likelihood
• Scalable. Seems to work in practice but somewhat ad hoc
• Robust to missing data
MMVAE: Architecture

- Mixture of experts: Equitable distribution of power among experts

\[ q(\mathbf{z}|\mathbf{x}_1, \mathbf{x}_2) = \frac{q_1(\mathbf{z}|\mathbf{x}_1) + q_2(\mathbf{z}|\mathbf{x}_2)}{2} \]

- Missing expert: \[ q(\mathbf{z}|\mathbf{x}_i) = 0 \]

Shi et. al, 2019
MMVAE: Loss function

\[ \mathcal{L}_{ELBO}(x_{1:M}) = \mathbb{E}_{z \sim q_{\phi}(z | x_{1:M})} \left[ \log \frac{p_{\theta}(z, x_{1:M})}{q_{\phi}(z | x_{1:M})} \right] \]

Importance weighted autoencoder:

\[ \mathcal{L}_{IWAE}(x_{1:M}) = \mathbb{E}_{z_{1:K} \sim q_{\phi}(z | x_{1:M})} \left[ \log \sum_{k=1}^{K} \frac{1}{K} p_{\theta}(z_{k}^{K} | x_{1:M}) \right] \]

\[ \mathcal{L}_{MoE}^{IWAE} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{z_{1:K} \sim q_{\phi}(z | x_{1:M})} \left[ \log \sum_{k=1}^{K} \frac{1}{K} p_{\theta}(z_{m}^{K} | x_{1:M}) \right] \]
Wish-list for multi-modal generative model

(a) Latent Factorisation

(b) Joint Generation

(c) Cross Generation

(d) Synergy
Reconstruction and Cross Generation

Cross Generation
Joint Generation

\[ z \]

\( \hat{x}_1 \) \( \theta_1 \)

\( \hat{x}_2 \) \( \theta_2 \)

Joint Generation

Same latent
Latent Factorization

shared information

private information of modality 1

Latent Factorisation

Latent traversal
Synergy

(d) Synergy

Log likelihoods

<table>
<thead>
<tr>
<th></th>
<th>$\log p(x_m \mid x_m, x_n)$</th>
<th>$\log p(x_m \mid x_n)$</th>
<th>$\log p(x_m \mid x_m)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m = MNIST$</td>
<td>868.76</td>
<td>628.31</td>
<td>868.37</td>
</tr>
<tr>
<td>$n = SVHN$</td>
<td>3441.01</td>
<td>2337.56</td>
<td>3441.01</td>
</tr>
</tbody>
</table>

Joint marginal likelihood $\geq$ Single marginal likelihood
Newer developments

- MoPoE-VAE
  \[ q_{\phi}(z|x_1, x_2) = \frac{q_{\phi_1}(z|x_1) + q_{\phi_2}(z|x_2)}{2} + p(z) \prod_{i=1}^{2} q_{\phi_i}(z|x_i) \]
  Sutter et. al. 2021

- MEME

  Joy et. al. 2021